Recognizing Induced Emotions of Movie Audiences: Are Induced and Perceived Emotions the Same?

Recognizing Induced Emotions of Movie Audiences: Are Induced and Perceived Emotions the Same?

Leimin Tian*, Michal Muszynski†, Catherine Lai*, Johanna D. Moore*,
Theodoros Kostoulas‡§, Patrizia Lombardo¶, Thierry Pun§, and Guillaume Chanel§
*School of Informatics, the University of Edinburgh
Emails: s1219694@sms.ed.ac.uk, clai@inf.ed.ac.uk, J.Moore@ed.ac.uk
†Computer Vision and Multimedia Laboratory, University of Geneva
Email: michal.muszynski@unige.ch
‡Faculty of Science and Technology, Bournemouth University, UK
Email: tkostoulas@bournemouth.ac.uk
§Computer Vision and Multimedia Laboratory & Swiss Center for Affective Sciences, University of Geneva
Emails: thierry.pun@unige.ch, guillaume.chanel@unige.ch
¶Department of Modern French & Swiss Center for Affective Sciences, University of Geneva
Email: patrizia.lombardo@unige.ch

***Abstract*—Predicting the emotional response of movie audiences to affective movie content is a challenging task in affective computing. Previous work has focused on using audiovisual movie content to predict movie induced emotions. However, the relationship between the audience's perceptions of the affective movie content (perceived emotions) and the emotions evoked in the audience (induced emotions) remains unexplored. In this work, we address the relationship between perceived and induced emotions in movies, and identify features and modelling approaches effective for predicting movie induced emotions. First, we extend the LIRIS-ACCEDE database by annotating perceived emotions in a crowd-sourced manner, and find that perceived and induced emotions are not always consistent. Second, we show that dialogue events and aesthetic highlights are effective predictors of movie induced emotions. In addition to movie based features, we also study physiological and behavioural measurements of audiences. Our experiments show that induced emotion recognition can benefit from including temporal context and from including multimodal information. Our study bridges the gap between affective content analysis and induced emotion prediction.**

1. Introduction

Recently, increased attention has been paid to recognizing emotions in spectators induced by affective content due to potential applications, such as emotion-based content delivery [1] or video indexing and summarization [2]. However, recognizing the emotions induced by affective content remains a challenging task, with only weak to moderate correlations achieved between automatic predictions and human annotations [3]. This limits the efficacy of affective content analysis in related applications.

When selecting stimuli to induce emotions, it is often assumed that emotions conveyed by the affective content (**perceived emotions** of the stimuli) are consistent with emotions evoked in the spectators (**induced emotions**). Perceived and induced emotions are often not distinguished in affective research. In fact, only a handful of previous studies have addressed the differences between the perceived emotions of affective content and the induced emotions of the spectator, mainly in music emotion research [4]. However, an empirical study has shown that emotions perceived from music are not always consistent with the emotional responses evoked in the audience [5]. This suggests the necessity of distinguishing perceived and induced emotions of movie audiences.

In comparison to music, movies convey complex information through multiple modalities. We identify three perspectives of emotions in movies: the audience's perspective, the actor's perspective, and the director's perspective. Movie audiences interpret the movie content and perceive the emotions it conveys (the perceived emotions). This then induces emotional responses which the audience feels (the induced emotions). Movie actors express emotions based on their interpretation of the script and may experience emotions themselves during acting (the expressed emotions). Movie directors create scripts with expectations of what emotions they intend the movie to induce in the audiences (the intended emotions). In this work we focus on the audience perspective of movie emotions. It has been argued that the perceived emotions of a movie can influence the induced emotional response of the audience by evoking empathy, which suggests a positive correlation between perceived and induced emotions [6]. However, Baveye et. al [7] argued that emotions intended by the directors may not always be consistent with emotions induced in the audience, although they did not discuss perceived emotions. To the best of

978-1-5386-0563-9/17/$31.00

our knowledge, there has been no previous work formally addressing the relationship between perceived and induced emotions of movie audiences. Therefore, we are motivated to bridge this gap by performing statistical analyses on emotions perceived from movie content and emotions induced in movie audiences. This will provide a foundation for understanding how affective content induces emotions in audiences, and how to use movie content information to predict movie induced emotions.

We choose the recently collected LIRIS-ACCEDE database [8] which contains continuous arousal-valence annotations of emotions induced in movie audiences for conducting experiments. This database has been widely used in state-of-the-art studies on movie induced emotions, including benchmark challenges such as MediaEval2016 [3]. We collect crowd-sourced annotations on perceived arousal, valence, and power of the movie dialogue to study the relationship between perceived and induced emotions.

State-of-the-art studies on recognizing induced emotions have focused on extracting features from the audiovisual content of the stimuli. However, lexical content, such as movie dialogue or lyrics of songs may also influence the emotional response of the audience. For example, movie dialogue has been shown to be effective for recognizing violence in movies [9]. Moreover, cues of perceived emotions in movies may be used for the recognition of induced emotions as well. Thus, we add manual transcripts of the LIRIS-ACCEDE movies, as well as expert annotations of DISfluency and Non-verbal Vocalisations (DIS-NV) in dialogues [10] and aesthetic highlights [11]. In addition, as a comparison with movie based features, we extract physiological and behavioural features based on signals collected from wearable sensors attached to the audience [12]. Beyond feature predictiveness, how the features are modelled also influences recognition performance. Thus, we study the impact of temporal context (history) on the recognition model and different fusion strategies for combining multimodal information (i.e., audio, visual, and lexical movie content, DIS-NV, aesthetic highlights, and physiological and behavioural measurements of movie audiences).

This work addresses the following research questions:

- Are perceived emotions of the movie content and induced emotions of the movie audience consistent?
- How can we improve performance when predicting movie induced emotions?
  - Do features beyond the audiovisual movie content contribute to recognizing induced emotions?
  - Does the recognition model benefit from including history and multimodal information?

The rest of this paper is arranged as: Section 2 reviews current affective content analysis studies. Section 3 contains protocols for collecting the extended LIRIS-ACCEDE annotation. Section 4 contains our analysis of the relationship between perceived and induced emotions of movie audiences. Section 5 presents unimodal and multimodal experiments on predicting movie induced emotions. The conclusion and future directions are provided in Section 6.

2. Related Work

The field of affective content analysis studies the relationship between information conveyed by the stimuli and emotional responses it evoked in the spectator. It remains a challenging task where only limited performance has been achieved for predicting induced emotions [7]. Here we briefly review the state-of-the-art of affective content analysis. First, we investigate previous work on the relationship between perceived and induced emotions to identify what has been missing. Second, we investigate previous work on induced emotion recognition on the LIRIS-ACCEDE database to identify limitations that we can improve on.

2.1. Perceived vs. Induced Emotions

When experiencing affective content, such as listening to music or watching a movie, we perceive emotions conveyed by the affective content from characteristics of the stimuli, such as tempi and pitch of music [5]. On the contrary, induced emotions of a spectator evoked by the stimuli are related to personal experience and individual preferences [13]. For example, a song perceived as happy reportedly induced stronger depression in a subject who is already in a depressed mood in [5]. Moreover, previous work indicates that perceived emotions are more objective than induced emotions [14], and annotators typically have stronger agreement over perceived emotions than induced emotions [15]. Previous work on affective content analysis does not always distinguish between perceived and induced emotions. Although consistencies between perceived and induced emotions have been found [16], music emotion research has identified fundamental differences between perceived and induced emotions (e.g. [5] and [17]). Previous work has also suggested that induced emotions can have more intensive arousal and less intensive valence ratings compared to perceived emotions of the same stimuli [4].

Compared to music emotions, there has only been limited work studying the relationship between perceived and induced emotions of movie audiences. Hanjalic and Xu [18] hypothesized positive correlations between perceived and induced emotions of movie audiences because perceived emotions can be used to estimate a spectator's affective reactions. Benini et al. [19] found that the annotator agreement on movie emotion descriptions is stronger when movie video features are included in the emotion definition. This also suggests a relationship between movie content and induced emotions. However, to the best of our knowledge, there has been no previous work studying how perceived and induced emotions of movie audiences are related.

2.2. Previous Work on LIRIS-ACCEDE Database

The LIRIS-ACCEDE database was collected and released to provide resources for researchers to collaborate on affective content analysis [8]. Here we focus on the continuous subset of the LIRIS-ACCEDE database, which contains 30 full movies, totalling 442 minutes [20]. During

data collection, 10 participants watched each movie once and annotated continuous arousal and valence scores (value range [−1,1]) of the emotions they felt during watching (movie induced emotions). The means of scores given by the participants over each second of the movie were used as the gold-standard annotations. A follow-up study screened these 30 movies to another 13 participants wearing sensors and collected physiological and behavioural measurements of the audiences during the movie [21].

Previous work on the LIRIS-ACCEDE database predicted movie induced emotions with various regression models, such as Support Vector Regression (SVR) [22], Long Short-Term Memory Recurrent Neural Networks (LSTM) [23], and Convolutional Neural Networks [24]. The Pearson Correlation Coefficient (CC) is the most commonly reported evaluation metric. Mean Squared Error (MSE) is sometimes reported in addition (e.g. [8]). Only weak or moderate correlations have been achived in state-of-the-art studies,[1] which shows that recognizing induced emotions of movie audiences is a challenging task. Note that different studies have different experiment protocols, such as data preprocessing and training-testing partitions. Thus, their results may not be directly comparable. Previous work has focused on using features extracted from audiovisual movie content (e.g. [25] and [26]). However, lexical information from the movie dialogue is overlooked, even though it has proved to be important in other emotion recognition studies [27]. Moreover, the usefulness of knowledge-inspired affective cues in movies, such as aesthetic highlights [21], has not been explored for predicting movie induced emotions.

Many previous studies examine unimodal models for induced emotion recognition (e.g. [28] and [29]). In fact, Baveye et al. [20] built a SVR model using only visual features and achieved best reported CC for this task. However, combining multimodal information has improved performance for a number of other emotion recognition tasks (e.g. [30]). Thus, we are motivated to study modality fusion strategies that may benefit induced emotion recognition. In addition, the LSTM model has low performance for predicting movie induced emotions [31],[2] yet it has achieved leading performance in various emotion recognition tasks due to its ability to model temporal context (e.g. [32]). Ma et al. [31] predict movie induced emotions at an interval of 10 seconds, which already contains temporal context. This may limit the gain when using a LSTM model to include more history. However, the suitable amount of history to include for predicting movie induced emotions is unclear.

3. Extended Annotations of LIRIS-ACCEDE

Here we provide two protocols for collecting the extended annotations of the continuous LIRIS-ACCEDE database. These include transcripts of movie dialogue with word timings and affective cue labels in Section 3.1, and perceived emotion annotations in Section 3.2. We choose 8 English movies listed in Table 1 which contain relatively more dialogue. Moreover, these movies are in the double-reality art form, where the lead characters switch between two worlds. This mirrors the activity of movie-watching where the reality and the movie world together create double-reality experience for the movie audience. Thus, the audiences may empathize more with the movie characters which is particularly interesting for understanding perceived and induced emotions. In total, we annotated 118 minutes of movies containing 870 utterances.

TABLE 1. STATISTICS OF SELECTED LIRIS-ACCEDE MOVIES

Movie	Genre	Utterance Count	Mean Sent. Duration (s)
After the Rain	Drama	77	3.000
First Bite	Romance	54	2.056
Nuclear Family	Comedy	147	2.694
Payload	Adventure	121	2.488
Spaceman	Adventure	133	2.489
Superhero	Drama	161	2.832
Tears of Steel	Adventure	79	2.165
The Secret Number	Drama	98	2.724

3.1. Transcription and Affective Cue Annotation

The movie transcription and affective cue annotation was conducted by two expert annotators. To increase the annotation speed, audio recordings of the movie were first passed through the IBM Watson Speech to Text service[3], which provides automatic speech transcription with word timings. This auto-generated transcript was then manually corrected and annotated by the annotators in parallel, each annotating five movies. To evaluate the annotation agreement, *First Bite* and *Spaceman* were annotated by both annotators and we compute the Normalized Damerau-Levenshtein (NDL) distances [33] of their transcripts, as well as the Pearson Correlation Coefficient (CC) of the word timings.

NDL distance is a widely used measurement of the distance between two strings. It is computed as the minimum number of operations required to transform one string to the other, divided by the length of the longer string of the pair. NDL distance of 0 indicates that the two strings are identical. Thus values closer to 0 show stronger annotation agreement. We find that 94.8% of the words transcribed are identical for the two annotators, with average NDL distance of 0.049. Considering the average length of words is 4 characters in the compared transcript, an average NDL distance of 0.049 means for every five words there is less than one character difference. CC for the word and utterance timings of the transcript is reported in Table 2. As we can see, the utterance and word timings annotated by the two annotators are strongly correlated. This verifies that the two annotators strongly agreed on movie transcription.

The same annotators also annotated the following types of Disfluency and Non-verbal Vocalisation (DIS-NV) in movie dialogue: filled pauses (e.g., “eh” or “hmm”), fillers

1. The best reported CC for arousal is 0.337, for valence is 0.296 [20]
2. CC for arousal is 0.054, for valence is 0.017 [31]
3. https://www.ibm.com/watson/developercloud/speech-to-text.html

TABLE 2. MOVIE TRANSCRIPT AND DIS-NV LABEL AGREEMENT (CC)

Labels	Start Time	End Time
Utterance	0.998	0.998
Word	0.999	0.999
General lexicon	0.989	0.989
Filled pause	0.625	0.625
Filler	0.920	0.920
Stutter	0.916	0.916
Laughter	0.635	0.635
Audible breath	0.766	0.764

(verbal filled pauses), stutters, laughter, and audible breath (remaining words are labelled as general lexicons). DIS-NVs were shown to be indicators of speaker emotions in spontaneous dialogue [10]. To evaluate annotation agreement, we divide the annotations into six subsets based on the DIS-NV labels and compute CC of the word timings in each subset. As shown in Table 2, although the annotation agreement on DIS-NV labels is lower compared to movie transcription, the annotations remain strongly correlated.

3.2. Annotating Perceived Movie Emotions

Emotion annotation is more subjective compared to movie transcription. Previous work has suggested that to achieve reliable emotion annotations, it is desirable to have more than 6 annotators [34]. To collect a large amount of annotations in a time and cost efficient manner, we annotate perceived emotions of movie audience using Amazon Mechanical Turk,[4] a crowd-sourced annotation platform. We segment movies into utterance clips using manual transcription of utterance timings and collect at least 10 annotations from different annotators for each clip. The annotators were instructed to rate the emotions expressed by movie characters on arousal (A), power (P), and valence (V) dimensions with 1 to 9 integer scores. We also provide explanations of each emotion dimension and meaning of different scores. Each Human Intelligence Task (HIT) contains clips of 5 continuous utterances from the same movie in their original order to provide movie context to the annotators. Each utterance appears at each of the five video windows in different HITs to reduce bias. The HITs are in random order and we kept track of previous annotators of each movie to prevent a utterance being annotated by the same annotator more than once. Annotators were only allowed to annotate a clip after it finished playing, and could only submit after annotating all clips. We published 1809 HITs and 129 annotators with various cultural and educational backgrounds participated. An example of the annotation interface is shown in Figure 1.

The 1 to 9 scores collected from the crowd-sourced annotation are normalized to [−1,1] to be consistent with the induced emotion annotation. We compute means of the annotations collected on each utterance of the movie dialogue as the perceived emotion annotation, resulting in utterance-level arousal, power, and valence annotations of perceived emotion of movie audiences.

4. https://requester.mturk.com/

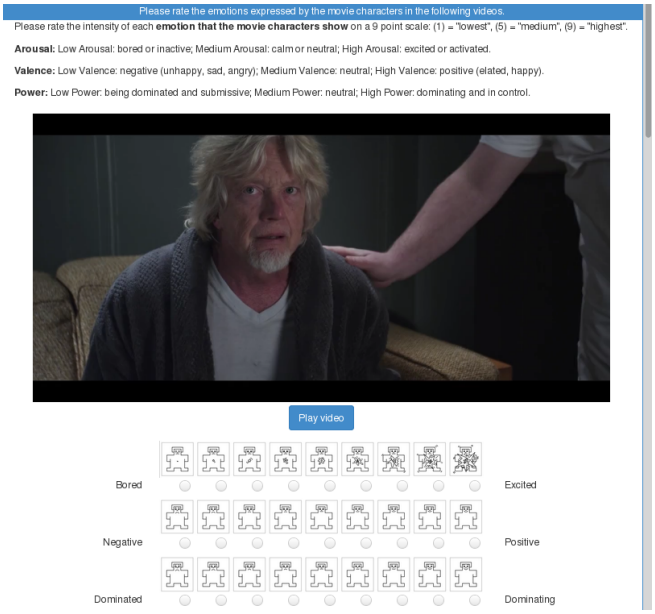

Figure 1. An example of Amazon Mechanical Turk annotation interface

4. Perceived and Induced Emotions

Here we address our first research question on the relationship between perceived and induced emotions of movie audiences. Note that the induced emotions are annotated at each second, while the perceived emotions are annotated at utterance-level and are generally longer than one second. Thus, we align the annotations by computing mean values of induced arousal-valence scores over each movie utterance as the utterance-level induced emotion annotation. We then calculate the CC between perceived and induced emotions for each movie independently. We use the fixed-effects model [35] to analize the CC between perceived and induced emotions. Thus, we computed weighted average of CC over all 8 movies and reported the resutls in Table 3. In the first row, "Per" is perceived emotions, "Ind" is induced emotions. To evaluate the practical significance of CC, following Cohen's model [36], we interpret absolute CC values at around 0.1, 0.3, and 0.5 as reflecting the effect size of small, medium, and large magnitude respectively (coloured as yellow, blue, and red in Table 3).

As we can see, perceived arousal, power, and valence are highly positively correlated with each other, while induced arousal and valence are moderately negatively correlated with each other. This may be related to perceived emotion annotation being a more objective task. The negative correlation between induced arousal and valence is consistent with previous work which found CC of −0.185 between crowd-sourced annotations of induced arousal and valence collected for nearly 14,000 English lemmas [37]. This suggests that induced negative emotions may have stronger arousal than induced positive emotions. However, no definitive conclusions can be made because of the small absolute CC value. Induced valence and perceived emotions have moderately positive correlations, while induced arousal and

TABLE 3. CC BETWEEN PERCEIVED AND INDUCED EMOTIONS

Emotion	Per-A	Per-V	Per-P	Ind-A
Per-V	0.538	#	#	#
Per-P	0.652	0.471	#	#
Ind-A	-0.095	-0.366	-0.170	#
Ind-V	0.243	0.345	0.307	-0.388

perceived emotions are weakly or moderately negatively correlated. In particular, perceived arousal and induced arousal are only weakly negatively correlated. This inconsistency between perceived and induced emotions indicates fundamental differences between perceived and induced emotions in movies. Emotion induction is a complex process. Various factors other than the emotions the movie content conveys can influence what emotional response is evoked in the audiences. The assumption that perceived and induced emotions are consistent is not accurate and researchers need to take extra precaution when designing experiments for affective content analysis research.

5. Recognizing Induced Emotions

To address our second research question, we investigate effective features and modelling approaches for predicting movie induced emotions. The original arousal-valence annotations on the LIRIS-ACCEDE database are provided for each second of the movie. To include suitable amount of data for feature extraction, we use a 5 second sliding window with a 4 second overlap between neighbouring windows to compute all features. The average arousal-valence scores over each window are used as the gold-standard induced emotion annotations. We also remove the end credits of each movie because participants started to remove the wearable sensors at this point, which introduce outliers in the signals. This results in 7103 data instances in total.

We perform leave-one-movie-out cross-validation and report unweighted average of MSE and absolute CC. MSE and CC are the most commonly reported evaluation metrics in related work (see Section 2.2). Higher CC represents stronger correlation between the predictions and the annotations and lower MSE represents smaller absolute value error and are thus desired. We evaluate the significance of the performance differences with two-sample Wilcoxon tests with $p < 0.05$ being significant. We compare arousal and valence predictions for pairs of models that have the closest performance in each experiment (e.g. Visual vs. DIS-NV on Arousal in Table 4), and find that all of them are significantly different with $p << 0.0001$, except for Lexical vs. Highlight on valence in Table 4 which has $p = 0.424$.[5] Note that because of different data processing procedures, such as the use of the overlapping window, our results are not directly comparable with previous work.

5. Lexical vs. Highlight on arousal has $p << 0.0001$. The second closest pair on valence (Audio vs. DIS-NV) has $p << 0.0001$

5.1. Audience Reaction Based Features

To compare with movie based features, we include two audience reaction based feature sets, namely physiological features and behavioural features. The physiological and behavioural signals of the movie audience are filtered by a third order low-pass Butterworth filter with cut-off frequency at 0.3Hz before feature extraction. The 273 physiological features are statistics over the sliding window based on the original measurement of the electrodermal activity of the audience and its first and second derivatives [21]. The 273 behavioural features are statistics over the sliding window based on the original measurement of signals collected from acceleration sensors attached to the audience's hands and its first and second derivatives [38]. Note that these physiological and behavioural measurements are collected from a different group of participants than those whose induced emotions were annotated as the gold-standard we are predicting in the unimodal and multimodal experiments.

5.2. Movie Based Features

Similar to previous work (e.g. [31]), we extract features from the audiovisual movie content with OpenSMILE [39]. For each sliding window, we extract 1582 InterSpeech2010 Paralinguistic Challenge Low-Level Descriptor audio features [40] and 1793 visual features. The later are histograms of Local Binary Pattern, HSV (hue, saturation, and value), and optical flow of each image region [41]. These are standard benchmark features used in various emotion recognition tasks [27]. To reduce feature dimensionality, we apply the ReliefF algorithm [42] and rank the individual effectiveness of features by performing regression with 20 nearest neighbours. We select the top 100 audio features and the top 100 visual features for arousal and valence respectively in order to maintain similar feature set sizes between different feature sets. We plan to investigate other feature engineering settings in future studies. We conduct ReliefF feature ranking on the remaining 22 movies of the continuous LIRIS-ACCEDE database outside the 8 movies we perform recognition experiments on. This allows us to incorporate in-domain knowledge and avoid including test data during feature selection.

Besides the data-driven audiovisual features, we extract three knowledge-inspired feature sets in addition. These include lexical features computed from the movie transcript, DIS-NV features, and aesthetic movie highlights.

The lexical features are based on crowd-sourced annotations of arousal, power, and valence ratings of 13,915 English lemmas [37] (i.e., the CSA features of Tian et al. [30]). We remove stop words from the movie transcript and lemmatize the remaining words. To compute the feature values, we search for the lemmas in each sliding window in the dictionary of [37]. Each dictionary entry contains 63 statistics calculated over the collected arousal, power, and valence ratings (21 for each emotion dimension). Sums of each of the 63 statistics for all the lemmas in the sliding window are returned as the 63 lexical features.

The six DIS-NV features are computed as the total duration of each type of DIS-NV (see Section 3.1) in each sliding window divided by the window length (5s). The lexical and DIS-NV features were shown to be effective predictors of speaker emotions in spontaneous dialogue [30].

The aesthetic movie highlights correspond to critical movie moments defined by experts in terms of art form and content [11]. They are knowledge-inspired cues and are more abstract than the audiovisual movie content. We record occurrences of six aesthetic highlights in each window:

- Spectacular: technical choices and special effects
- Subtle: camera use, lighting, and music
- Character: emotions and responses to dramatic events
- Dialogue: clarifying motivation and showing tension
- Theme: unusual close-up and theme development
- Any type of highlight above has occurred

The knowledge-inspired features are more sparse than the audiovisual features. These dialogue cues and highlights are infrequent events in the movie. Thus, majority of the knowledge-inspired feature values are zero vectors.

5.3. Recognition Models

We build LSTM models with the Keras library [43] for regression. RMSprop with a learning rate of 0.0001 and the MSE evaluation metric is used for training. All LSTM models have three hidden layers (number of neurons: $h_1 = 64$, $h_2 = 32$, $h_3 = 16$). To prevent over-fitting, we use 0.5 drop-out rate in h_1 and set the maximum training iteration to 50 epochs with an early stopping tolerance of 10 epochs. This LSTM structure has been shown to be robust for emotion recognition in previous work [30].

For multimodal experiments, we test three fusion strategies: Feature-Level (FL) fusion (also known as "early fusion"), Decision-Level (DL) fusion (also known as "late fusion"), and HierarchicaL (HL) fusion [30]. In FL fusion, all features are concatenated before input to the recognition model. In DL fusion, unimodal recognition models for each feature set are built and their outputs are used in a decision-making module. The HL fusion strategy uses different features in different levels of its hierarchy: noisy features are incorporated in lower levels, while more abstract features are incorporated in higher levels. In our multimodal models, for FL fusion, all features are used at the input layer of the LSTM model. For DL fusion, predictions of unimodal LSTM models are input to another LSTM model. For HL fusion, input neurons of low-level features are connected to h_1, while input neurons of high-level features are connected to h_2 directly. We build multimodal models combining all features, as well as multimodal models using only movie based features. For the HL model using all features, the physiological and behavioural features are used at the higher layer because they are measurements of audience's reactions. For the HL model using movie based features, the audiovisual features are used at the higher layer because we include in-domain knowledge during feature selection.

5.4. Results and Discussion

Here we discuss our experiments on predicting movie induced emotions. In unimodal experiments, we first study the influence of temporal context by building LSTM models with different time steps, then compare performance of predicting induced emotions with different features. In multimodal experiments, we study the gain of different fusion strategies for combining multimodal information.

5.4.1. Influence of history on induced emotion.

The original induced emotion annotation provided by the continuous LIRIS-ACCEDE database is at every single second, where the average absolute difference between adjacent emotion annotations of arousal is 0.006 and of valence is 0.005. This is extremely small considering the annotation value range is [−1,1]. Previous work has shown that human emotions are context dependent and typically do not change rapidly over a small time interval [27]. However, the suitable amount of temporal context for predicting movie induced emotions remains unknown.

We attempt to identify suitable amount of history for predicting induced emotions by testing LSTM model using physiological features with different time steps. We use physiological features because they are direct representatives of the audience's induced responses [38]. Our experiments show that including features for the past 3 time steps gives better recognition performance than shorter or longer time steps. Thus, later LSTM models in this work all use a time step of 3. Recall that our feature vectors are extracted over a 5 second sliding window with 4 seconds overlap. With 3 history feature vectors the model will have 8 seconds of temporal context (including the current window).

5.4.2. Unimodal induced emotion recognition.

Results of our unimodal induced emotion recognition experiments are shown in Table 4. Numbers in bold indicate the best performance for the experiment. As we can see, the physiological features achieved the best CC on predicting induced valence, even though they are based on measurements of a different audience than the audience whose induced emotions are being predicted. This indicates that people share similarities in how and what emotions are induced by the same movie. The behavioural features are less predictive than the physiological features. This indicates that hand movements of the audience may be caused by various factors besides induced emotions, and contain more noise compared to the electrodermal measure. The audio features achieved the best CC on predicting induced arousal. This suggests that including in-domain knowledge can benefit induced emotion recognition. Knowledge-inspired features based on affective cues achieved better MSE on predicting induced arousal and valence than data-driven features based on audiovisual movie content. This different behaviour of CC and MSE shows that an evaluation metric combining correlation and error may be better for evaluating induced emotion recognition performance. For example, the concordance correlation coefficient [44]. Features based on audiovisual movie

TABLE 4. UNIMODAL INDUCED EMOTION RECOGNITION (MSE&CC)

Features	A-mse	A-cc	V-mse	V-cc
Audience Reaction Features				
Physiological	**0.047**	**0.190**	0.066	**0.432**
Behavioural	0.049	0.183	**0.064**	0.129
Movie Based Features				
Audio	0.054	**0.218**	0.069	0.134
Visual	0.060	0.126	0.090	**0.152**
Lexical	0.050	0.085	0.071	0.060
DIS-NV	0.049	0.124	**0.069**	0.115
Highlight	**0.049**	0.153	0.070	0.056

content being predictive of induced emotions is consistent with our findings in Section 4 that perceived emotions are only moderately correlated with induced emotions. Emotion induction is a complex process and factors other than perceived emotions also influence induced emotions.

5.4.3. Multimodal induced emotion recognition.

Table 5 contains our multimodal induced emotion recognition results. We build multimodal models both with all features and with only movie based features. For multimodal models using all features, the FL model has the best CC for predicting arousal, while the HL model has the best CC for predicting valence. Recall that the physiological and behavioural features are used at a higher layer than other features in HL fusion. In Table 4, the audio features have the best CC for predicting arousal, while the physiological features have the best CC for predicting valence. The audio features have larger influence in FL fusion than in HL or DL fusion, resulting in better CC for predicting arousal using FL fusion. The DL model has the best MSE. This may be related to DL fusion not being influenced by feature dimension. Thus, the DL model benefits more from the smaller DIS-NV and Highlight feature sets which have good MSE performance. The multimodal models outperform the unimodal models on predicting arousal, but not valence. This may be caused by a lack of training data.

Multimodal models using only movie based features have significantly worse performance than those using all features. For multimodal models using movie based features, the DL model has the best performance, except for CC of arousal. This is because the audio features, which have best unimodal CC for predicting induced arousal, have larger influence in FL fusion than in DL fusion. Unlike previous work [30], HL fusion does not outperform FL or DL fusion here. The reason may be that we extract all features using overlapping windows and we reduce noise in the audiovisual features by feature selection. Thus, the difference between movie based features in terms of abstraction level or time scale is not as large as [30], limiting the gain of HL fusion. Similar to multimodal models using all features, multimodal models using movie based features outperform unimodal models except for CC of arousal.

Our experiments indicate that for predicting movie induced emotions, performance improvements can be achieved by including temporal context, and by incorporating abstract affective cues in addition to the audiovisual movie content.

TABLE 5. MULTIMODAL INDUCED EMOTION RECOGNITION

Model	A-mse	A-cc	V-mse	V-cc
Using All Features				
FL	0.057	**0.271**	0.071	0.107
DL	**0.044**	0.189	**0.070**	0.163
HL	0.074	0.159	0.095	**0.227**
Using Movie Based Features				
FL	0.054	**0.216**	0.069	0.118
DL	**0.044**	0.182	**0.057**	**0.178**
HL	0.073	0.157	0.075	0.083

6. Conclusion

This work bridges the gap between perceived and induced emotions of movie audiences and serves as a reference for future affective content analysis studies. We extend annotations on the continuous LIRIS-ACCEDE database and find that perceived and induced emotions of movie audiences are not always positively correlated. When selecting stimuli for emotion induction, there is more to be considered than simply assuming that the perceived emotions of the stimuli will be consistent with the emotions induced in spectators. To expand our understanding of perceived and induced emotions, we plan to investigate using perceived emotions to predict induced emotions. Besides perceived and induced emotions of the movie audiences, we would also like to include other perspectives of movie emotions to study the complete relationship between the three perspectives of movie emotions. In addition, we plan on conducting further investigations on how emotions and affective cues differ in different movie genres. The study between different perspectives of movie emotions may contribute to the movie art research as well and help film directors design affective contents aligning intended and induced emotions better.

Our unimodal and multimodal experiments on predicting movie induced emotions indicate that it is beneficial to include temporal context and to combine knowledge-inspired affective cues with audiovisual movie content. Our results show that the small amount of labelled data available for affective content analysis can limit performance significantly. Inspired by audiovisual features benefiting from including in-domain knowledge, we will study the gain of applying transfer learning for predicting movie induced emotions. Note that our induced emotion recognition experiments here are preliminary. Improved performance may be achieved by optimizing feature representations and model structures, which we will investigate in our future studies.

Acknowledgments

Leimin Tian is funded by School of Informatics, the University of Edinburgh. Michal Muszynski is funded by Computer Vision and Multimedia Laboratory, the University of Geneva. This work is partially supported by grants from the Swiss Center for Affective Sciences and the Swiss National Science Foundation. We want to thank Mohammad Soleymani, Anna Aljanaki, and Soheil Rayatdoost for their generous help on Amazon Mechanical Turk experiments.

References

[1] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized tv," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 90–100, 2006.

[2] S. Arifin and P. Y. Cheung, "Affective level video segmentation by utilizing the pleasure-arousal-dominance information," *IEEE Transactions on Multimedia*, vol. 10, no. 7, pp. 1325–1341, 2008.

[3] E. Dellandréa, L. Chen, Y. Baveye, M. Sjöberg, C. Chamaret, and E. Lyon, "The mediaeval 2016 emotional impact of movies task," in *MediaEval2016*, 2016.

[4] K. Kallinen and N. Ravaja, "Emotion perceived and emotion felt: Same and different," *Musicae Scientiae*, vol. 10, no. 2, pp. 191–213, 2006.

[5] A. Gabrielsson, "Emotion perceived and emotion felt: Same or different?" *Musicae Scientiae*, vol. 5, no. 1_suppl, pp. 123–147, 2001.

[6] E. S.-H. Tan, "Film-induced affect as a witness emotion," *Poetics*, vol. 23, no. 1-2, pp. 7–32, 1995.

[7] Y. Baveye, C. Chamaret, E. Dellandréa, and L. Chen, "Affective video content analysis: A multidisciplinary insight," *IEEE Transactions on Affective Computing*, 2017.

[8] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43–55, 2015.

[9] G. Gninkoun and M. Soleymani, "Automatic violence scenes detection: A multi-modal approach," 2011.

[10] L. Tian, J. D. Moore, and C. Lai, "Emotion recognition in spontaneous and acted dialogues," in *ACII2015*. IEEE, 2015, pp. 698–704.

[11] T. Kostoulas, G. Chanel, M. Muszynski, P. Lombardo, and T. Pun, "Dynamic time warping of multimodal signals for detecting highlights in movies," in *INTERPERSONAL2015*. ACM, 2015, pp. 35–40.

[12] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun, "Affective ranking of movie scenes using physiological signals and content analysis," in *MS2008*. ACM, 2008, pp. 32–39.

[13] C. Plantinga, "Art moods and human moods in narrative cinema," *New Literary History*, vol. 43, no. 3, pp. 455–475, 2012.

[14] G. Matthews, D. M. Jones, and A. G. Chamberlain, "Refining the measurement of mood: The uwist mood adjective checklist," *British journal of psychology*, vol. 81, no. 1, pp. 17–42, 1990.

[15] Y. Song, S. Dixon, M. T. Pearce, and A. R. Halpern, "Perceived and induced emotion responses to popular music," *Music Perception: An Interdisciplinary Journal*, vol. 33, no. 4, pp. 472–492, 2016.

[16] K. Knautz and W. G. Stock, "Collective indexing of emotions in videos," *Journal of Documentation*, vol. 67, no. 6, pp. 975–994, 2011.

[17] J. Tarvainen, M. Sjöberg, S. Westman, J. Laaksonen, and P. Oittinen, "Content-based prediction of movie style, aesthetics, and affect: Data set and baseline experiments," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2085–2098, 2014.

[18] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE transactions on Multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

[19] S. Benini, L. Canini, and R. Leonardi, "A connotative space for supporting movie affective recommendation," *IEEE Transactions on Multimedia*, vol. 13, no. 6, pp. 1356–1370, 2011.

[20] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "Deep learning vs. kernel methods: Performance for emotion prediction in videos," in *ACII2015*. IEEE, 2015, pp. 77–83.

[21] T. Li, Y. Baveye, C. Chamaret, E. Dellandréa, and L. Chen, "Continuous arousal self-assessments validation using real-time physiological responses," in *ASM2015*. ACM, 2015, pp. 39–44.

[22] A. M. Andrew, "An introduction to support vector machines and other kernel-based learning methods by nello christianini and john shawe-taylor," 2000.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] D. C. Cireşan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Deep, big, simple neural nets for handwritten digit recognition," *Neural computation*, vol. 22, no. 12, pp. 3207–3220, 2010.

[25] T. Anastasia and H. Leontios, "Auth-sgp in mediaeval 2016 emotional impact of movies task," 2016.

[26] S. Chen and Q. Jin, "Ruc at mediaeval 2016 emotional impact of movies task: Fusion of multimodal features," 2016.

[27] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[28] A. Jan, Y. F. A. Gaus, F. Zhang, and H. Meng, "Bul in mediaeval 2016 emotional impact of movies task," 2016.

[29] Y. Liu, Z. Gu, Y. Zhang, and Y. Liu, "Mining emotional features of movies," 2016.

[30] L. Tian, J. Moore, and C. Lai, "Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features," *SLT2016*, 2016.

[31] Y. Ma, Z. Ye, and M. Xu, "Thu-hcsi at mediaeval 2016: Emotional impact of movies task," 2016.

[32] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in *AVEC2016*. ACM, 2016, pp. 97–104.

[33] G. V. Bard, "Spelling-error tolerant, order-independent pass-phrases via the damerau-levenshtein string-edit distance metric," in *ACSW2007*, vol. 68. Australian Computer Society, Inc., 2007, pp. 117–124.

[34] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[35] J. Sánchez-Meca and F. Marín-Martínez, "Meta-analysis in psychological research." *International Journal of Psychological Research*, vol. 3, no. 1, pp. 150–162, 2015.

[36] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*, 2nd ed. Routledge, 1988.

[37] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior research methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[38] T. Kostoulas, G. Chanel, M. Muszynski, P. Lombardo, and T. Pun, "Identifying aesthetic highlights in movies from clustering of physiological and behavioral signals," in *QoMEX2015*. IEEE, 2015, pp. 1–6.

[39] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the munich versatile and fast open-source audio feature extractor," in *ICMI2010*. ACM, 2010, pp. 1459–1462.

[40] B. W. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, S. S. Narayanan *et al.*, "The interspeech 2010 paralinguistic challenge." in *Interspeech*, vol. 2010, 2010, pp. 2795–2798.

[41] F. Eyben, F. Weninger, M. Wöllmer, and B. Schuller, "open-source media interpretation by large feature-space extraction," 2016.

[42] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," *Machine learning*, vol. 53, no. 1-2, pp. 23–69, 2003.

[43] F. Chollet, "Keras," https://github.com/fchollet/keras, 2015.

[44] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.